chinaXiv:202112.00002v1

深度学习损失函数地貌分析研究进展

梁若冰 [1,2]，刘波 [1*]，孙越泓 [3]

(1. 中国科学院 数学与系统科学研究院，北京 100190; 2. 中国科学院大学，北京 100049; 3. 南京师范大学 数学科学学院，南京 210046)

摘要 在机器学习和数学优化研究领域，深度学习优化问题易优性的数学解释极具挑战性. 损失函数存在高维、非凸、不光滑等特质性，然而也能通过梯度下降法搜索到全局最优值. 损失函数地貌分析成为揭示深度学习优化问题易优性本质的重要研究方向. 为促进可解释、可信的深度学习在更关键领域的应用，本文回顾了损失函数地貌特征（局部极小点的数量和空间分布、最优点之间的连通性、临界点的最优性）、梯度下降法收敛性、以及损失函数地貌可视化等方面的研究进展和挑战.

关键词 深度学习；损失函数；地貌分析

Loss landscape analysis for deep learning: A survey

LIANG Ruobing[1,2], LIU Bo[1*], SUN Yuehong[3]

(1. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. School of Mathematical Sciences, Nanjing Normal University, Nanjing 210046, China)

Abstract In the filed of machine learning and mathematical optimization, it is a challenge to mathematically explain optimality of loss function for deep learning. Loss function is high-dimensional, non-convex, and non-smooth. It was, however, observed that gradient descent could reach zero training loss of this highly non-convex function. Loss landscape analysis is critical to reveal reasons why deep networks are easily optimizable. We reviewed the advance on loss landscape analysis, such as landscape features (number and spatial distribution of local minima, connectivity between global optima, and global optimality of critical points), convergence of gradient descent, and visualization of loss landscape. This survey aimed to promote interpretable and reliable deep learning in critical applications.

Keywords deep learning; loss function; landscape analysis

1. 引言

深度学习通过由神经元、连接权、偏置值与激活函数组成的多层网络结构 [1]，逐层将低层特征表达转化为高层特征表达，借由最优或较优的输入到输出的映射，完成复杂的表示学习任务 [1, 2]. 比如，深度学习在识别肺癌时，利用监督学习、半监督学习甚至无监督学习来进行特征

收稿日期:
作者简介:. 梁若冰 (1997-)，女，博士研究生，研究方向：机器学习，Email: rbliang@amss.ac.cn; 孙越泓 (1972-)，女，副教授，研究方向：智能优化，Email: 05234@njnu.edu.cn; 通讯作者：刘波 (1979-)，男，副研究员，研究方向：组合优化，Email: bliu@amss.ac.cn.
资助项目: 中国科学院前沿科学重点研究计划 (QYZDB-SSW-SYS020)
Foundation items: Frontier Science Key Research Program, Chinese Academy of Sciences (QYZDB-SSW-SYS020)

chinaXiv:202112.00002v1

学习，借由较低层识别病灶边缘和边缘的组合，利用较高层进行概念识别 [3, 4]. 在克服了梯度消失 [5]、训练数据缺乏、计算能力弱等关键困难后，多种深度学习模型被成功应用于计算机视觉、商业智能、医学图像分析等，并在特定任务上的性能超过了有经验的人类专家 [6].

损失函数是以网络参数为因变量的非负函数，度量神经网络在表示学习中输出值与真实值的差异. 深度学习利用梯度下降等优化算法，调节网络参数，以损失函数最小化为优化目标，直至网络的输出值与真实值一致或接近，此时网络实现了从输入到输出的最优映射 [1].

损失函数存在高维、非凸和不光滑等困难性质，对优化理论和算法研究造成了挑战. 利用深层神经网络进行表示学习时，一个神秘的现象是：针对非凸的损失函数，即使采用随机初始化的一阶梯度下降法训练深度神经网络，也能使损失函数在训练集上收敛到全局最优值，即零训练损失 [7, 8]. 梯度下降法实现零训练损失的现象，赋予了深度网络良好的拟合学习能力. 然而，相关数学优化理论匮乏，诸多研究一直在探究这一有悖于优化理论的经验观察背后的原因. 其中，用于解释深度网络具备较强表示能力的过参数化 [9-11]和万能逼近定理 [12-14]，无法解释梯度下降法实现零训练损失的现象.

深度学习优化问题最优性的数学解释是极具挑战的研究. 自 2015 年起，机器学习、数学优化等领域的学者开始分析深度学习损失函数地貌特征，尝试给出深度学习优化问题易优性的数学解释. 经过五年多的发展，在损失函数地貌特征分析、梯度下降法收敛性分析、以及损失函数地貌可视化等三方面取得了长足的进展. 深度学习损失函数地貌分析已成为揭示深度学习优化问题本质、刻画最优解结构性质、分析优化算法收敛性的有效数学分析工具. 本文回顾了深度学习损失函数地貌分析研究进展，给出了面临的挑战，并展望了未来的研究方向.

本文安排如下：第二章是关于损失函数和梯度下降法的预备知识；第三章给出损失函数地貌特征分析研究进展；第四章给出梯度下降法收敛性分析进展；第五章给出损失函数地貌可视化研究进展；第六章给出挑战和进一步的研究.

2. 预备知识

2.1 损失函数

损失函数定义为神经网络在训练样本集 $D=\{x_i, y_i\}_{i=1}^n$ 上的误差，是网络参数 $\theta=(W_i,b_i)_{i=1}^L$ 的函数，其中 W_i 为第 i 层权重矩阵， b_i 为第 i 层节点偏置值向量. 公式(1)的 l_2 损失函数常用于连续型的表示学习任务，公式(2)的二分类交叉熵损失函数常用于离散型的表示学习任务.

$$L(y_i,\hat{y}_i)=\| y_i-\hat{y}_i \|^2 \tag{1}$$

$$L(y_i,\hat{y}_i)=(-y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)) \tag{2}$$

其中 $L(y_i,\hat{y}_i)$ 是样本 (x_i,y_i) 的损失函数， x_i 和 y_i 为实际输入和标签， $x_i\in R^{d_x}$ 为 d_x 维向量， $y_i\in R^{d_y}$ 为 d_y 维向量， $\hat{y}_i=f(x_i\,|\,\theta)$ 是在给定参数 θ 和实际输入 x_i 时的网络输出的标签预测值， $X\in R^{d_x\times n}$ 表示 n 个输入 x_i 作为列向量组成的数据矩阵， $Y\in R^{d_y\times n}$ 表示 n 个标签 y_i 作为列向量组成的标签矩阵， $\overline{Y}\in R^{d_y\times n}$ 表示 n 个网络输出 $\hat{y}_i$ 作为列向量组成的输出预测矩阵.

损失函数在数据生成分布上的期望最小化为

$$J(\theta)=E_{(x,y)\sim p_{data}}L(y,f(x\,|\,\theta)) \tag{3}$$

其中 p_{data} 是训练样本集 D 的分布.

在给定训练集上，只能计算经验风险 $\Re(\theta)$，即损失函数在训练集上的平均损失：

$$\Re(\theta)=\frac{1}{n}\sum_{i=1}^{n}L(y_i,\hat{y}_i)=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i\mid\theta)) \tag{4}$$

遵循经验风险最小化的原则，找到最优 $\theta*$，使得在训练集上的平均损失最小

$$\theta*=\arg\min_{\theta}\Re(\theta) \tag{5}$$

损失函数的参数维度非常高. LeNet-5 [15]约有 6 万参数，ResNet10 [16]约有 1 千万个的参数，AlexNet [17]约有 6 千万个的参数，而VGG16 [18]的参数量约为 1 亿.

损失函数是非凸函数. Kawaguchi [19]证明了深层神经网络的损失函数存在没有负特征值的临界点，损失函数不具备凸性. Dauphin等 [20]发现鞍点与局部极小点数量的比值随函数维度增加而指数级增加，鞍点处的损失函数值较大，且鞍点的邻域是较大的平坦区域，鞍点处具备较强的非凸性.

损失函数不是严格的光滑函数. 比如，激活函数 ReLU 在零点处不可导. 为了便于进行反向传播计算，定义该函数在零点处的导数值为零.

损失函数存在高维、非凸和不光滑等困难性质，对最优性的数学解释带来了极大的挑战.

2.2 梯度下降法

梯度下降法是常用的神经网络参数更新算法 [1]. 根据参数更新时需要的数据量不同，梯度下降法可以分为：批量梯度下降法(Batch Gradient Descent, BGD)、随机梯度下降法(Stochastic Gradient Descent, SGD)和小批量梯度下降法(Mini-batch Gradient Descent, MGD). BGD在获得所有样本的梯度后对网络参数进行更新. SGD则每次从训练集中随机选择一个样本进行学习，学习速度快，但较难收敛. MGD每次从训练集中选取 m 个样本($m\le n$)，降低了收敛震荡性，但需选取合适的样本数量.

3. 损失函数地貌特征分析

本节回顾了利用代理模型、矩阵分析、随机分析、微分几何等数学分析工具对深度学习损失函数地貌特征进行分析的进展，从局部极小点的数量与空间分布、全局最优点之间的连通性、临界点的最优性等方面理解损失函数地貌.

3.1 局部极小点的数量与空间分布

局部极小点的数量与空间分布有助于刻画问题的优化难度. 由于损失函数的高维特性，导致无法直接获得这类有益信息，一些研究致力于使用代理模型近似损失函数，以获得损失函数局部极小点的数量与空间分布等性质 [21].

Choromanska等 [22]发现了深层线性神经网络的损失函数与自旋玻璃模型的哈密顿量具有相似的性质. 高维损失函数的低指数临界点形成了一个分层的结构，临界点也是局部极小点，且位于以全局最优点为下界限定的一个有界区域中. 在该有界区域外，找到低指数临界点的概率随着损失函数维度的增加而指数阶降低. Becker和Zhang [23]在Choromanska等 [22]的基础上，利用随机矩阵和代数几何，发现了损失函数与球形旋转玻璃模型的哈密顿量同分布，将损失函数表示成为网络深度的函数. 当神经网络参数数量保持不变时，通过增加网络深度，损失函数的临界点数量减少，最优点在参数空间中更加聚集，从而使得损失函数更容易优化. Cooper [24]利用微分几何中的Sard's定理，证明了ReLU激活函数的全连接神经网络损失函数的临界点集为非空子流形，该子流形的维数是参数数量与样本个数的差值.

3.2 最优点之间的连通性

探索最优点之间的连通性或可达性，有助于刻画最优点在空间中的分布特征，并解释损失函数的易优性.

chinaXiv:202112.00002v1

Garipov等 [25]发现最优点之间可由一条简单曲线连接，取该曲线上任意一点作为新网络的参数，则新网络在训练集上的损失函数值与原网络几乎一致．这个发现为参数不同但具有相同表示学习能力的深度网络提供了一种新的几何学解释．Nguyen [26]利用了隐藏层输出线性独立、连通集合等性质，给出了分段线性激活函数的深层全连接网络凸损失函数的最优点集合的特征．若某隐藏层宽度大于训练集样本数量，且其后隐藏层宽度逐层递减，则各层权重矩阵均满秩，损失函数 L 的任意 α 水平子集 $\Gamma_\alpha = \{\theta \in \Omega \mid L(\theta) \le \alpha\}$ 连通，有且仅有一个包含了所有全局最优点的连通集.

3.3 临界点的最优性

临界点是损失函数的导数为零的点，包括鞍点、局部极小点和全局最优点．探究临界点集的性质，特别是临界点的最优性，有助于解释损失函数易优性.

3.3.1 深层线性神经网络

Kawaguchi [19]给出了线性激活函数的深层网络损失函数在临界点的Hessian矩阵半正定的充分条件，保证了临界点集合中不存在局部极小点．在 XX^T 和 XY^T 满秩，且 $YX^T(XX^T)^{-1}XY^T$ 的不同特征值个数等于标签向量维度时，临界点或为全局最优点或为鞍点，且任意鞍点的Hessian矩阵至少存在一个负特征值．Lu和Kawaguchi [27]进一步将Kawaguchi [19]中的条件放宽为 X 和 Y 行满秩，该证明依赖于矩阵的奇异值分解与矩阵奇异空间的连续性．Zhou和Liang [28]无需对Kawaguchi [19]中的网络参数和数据矩阵做任何假设，利用权重矩阵分解即可证明临界点集的局部极小点均为全局最优点．Yun等 [29]发现当权重矩阵的乘积为满秩矩阵时，l_2 损失函数的临界点只能为全局最优点或鞍点，并给出了区分鞍点和全局最优点的判据，即权重矩阵乘积为满秩矩阵的集合内的临界点为全局最优点，集合外的临界点则为鞍点．针对深层线性残差神经网络，Hardt和Ma [30]基于残差网络的恒等映射的特殊结构，证明了当网络权重矩阵的谱范数一致地小时，所有临界点都是全局最优点.

3.3.2 深层非线性神经网络

部分研究关注不同的非线性激活函数，比如，ReLU 函数、解析激活函数、光滑激活函数等对临界点最优性的影响.

针对ReLU激活函数的非线性深层神经网络，Kawaguchi [19]采用了Choromanska等 [22]中的部分假设，引入服从伯努利分布的随机向量，将原输出预测值表示为数据矩阵、权重矩阵和随机向量的乘积，进而构建了损失函数的期望，借助期望的平均特性抵消激活函数的非线性，将深层线性神经网络的结论推广到了深层非线性网络，即临界点或为全局最优点或为鞍点.

针对解析激活函数与 l_2 损失函数的全连接深层神经网络，Nguyen和Hein [31]讨论了临界点为全局最优点的充分条件．利用满秩矩阵和解析函数的性质，证明了只要某一隐藏层宽度大于输入数据维数，权重矩阵从第 k 层开始是行满秩矩阵，且从第 k 层开始网络节点数逐级减少，则非退化的临界点就是全局最优点，且不存在低秩的局部极小点．Nguyen和Hein [32]进一步研究了卷积神经网络中的临界点最优性．将卷积层运算重构为全连接层计算形式，发现了在所有满足使第 k 层输出线性独立，$k+2$ 层到 l 层的权重矩阵满秩的网络参数集合中，存在无数个临界点，这些临界点均为全局最优点.

针对光滑激活函数的非线性神经网络，Yun等 [29]定义了网络各层的函数空间，将损失函数定义为网络各层映射的函数．临界点定义为使损失函数对任意映射的Fréchet导数为零的点．该研究给出了函数空间中临界点为全局最优点的充分条件，但该结论无法推广到参数空间．函数空间中不存在无法下降的次优点，任意次优点在函数空间中都存在下降方向，然而将次优点对应到参数空间时，其在函数空间的下降方向可能与其在参数空间正交，因而该次优点可能对应于参数空间的局部极小点.

chinaXiv:202112.00002v1

表 1 损失函数地貌特征分析研究

作者	时间	网络类型	研究方法	结论及适应性	局限性
Choromanska 等 [22]	2015	ReLU 激活函数的二分类全连接深层神经网络	发现了深层线性神经网络的损失函数和自旋玻璃模型的哈密顿量具有相似的性质	临界点位于以全局最优点为下界限定的有界区域中	输入数据需要相互独立
Kawaguchi [19]	2016	深层线性神经网络	给出了损失函数在临界点处 Hessian 矩阵半正定的充分条件	任意局部极小点都是全局最优点	输入数据需要相互独立
		ReLU 激活函数的非线性深层神经网络	在损失函数中引入随机变量，借助期望的平均特性抵消激活函数的非线性，将深层线性网络的结论推广到了深层非线性网络	任意局部极小点都是全局最优点	随机变量、数据矩阵等需要满足前提假设
Hardt和Ma [30]	2016	线性残差神经网络	给出了损失函数对权重矩阵的偏导数满足的不等式	证明了当权重矩阵的谱范数一致地小时，所有临界点都是全局最优点	结论适用于线性残差神经网络
Lu 和 Kawaguchi [27]	2017	深层线性神经网络	利用矩阵奇异值分解和奇异空间连续性将深层网络归约为浅层网络	X 和 Y 行满秩时，损失函数的局部极小点均为全局最优点	结论仅适用于深层线性网络
Zhou 和 Liang [28]	2017	深层线性神经网络	利用权重矩阵分解	临界点集的局部极小点均为全局最优点	结论仅适用于深层线性网络和浅层非线性网络
Nguyen 和 Hein [31]	2017	解析激活函数的全连接深层神经网络	利用满秩矩阵和解析函数的性质，给出了临界点为全局最优点的充分条件	非退化的临界点都是全局最优点，且不存在低秩的局部极小点	需要输入数据线性独立，且激活函数为解析函数
Yun等 [29]	2017	深层线性神经网络	给出了损失函数值对权重矩阵的偏导数满足的不等式	权重矩阵乘积为满秩矩阵的集合内的临界点为全局最优点，集合外的临界点为鞍点	需要满足权重矩阵的乘积为满秩矩阵
		光滑激活函数的非线性神经网络	临界点为使得损失函数对任意映射的 Fréchet 导数为零的点	函数空间中临界点为全局最优点	结论无法推广到参数空间
Nguyen 和 Hein [32]	2018	卷积神经网络	将卷积层运算重构为全连接层计算，研究临界点的最优性	临界点都是全局最优点，且有无穷多个	激活函数为解析函数，且没有讨论存在池化层的卷积神经网络
Garipov等 [25]	2018	深层神经网络	最优点之间的连通性	发现可以找到一条简单曲线来连接最优点. 取该曲线上任意一点	没有给出损失函数地貌的全面特征

chinaXiv:202112.00002v1

				作为新网络的参数，则新网络在训练集上的损失函数值与原网络几乎一致	
Cooper [24]	2018	ReLU 激活函数的全连接神经网络	利用微分几何中的 Sard’s 定理分析局部极小点的数量和空间分布	损失函数零点集为非空的子流形，维数为参数个数与样本数量的差值	没有讨论子流形不为空集的更广泛条件
Nguyen [26]	2019	分段线性激活函数的深层全连接网络	利用了足够宽的隐藏层输出线性独立与连通集合等性质对最优点连通性进行研究	全局最优点都在一个连通集中，没有严格的局部极小点	水平集的连通性无法保证梯度下降算法的收敛性，仅直观地给出损失函数地貌特性
Becker和Zhang [23]	2020	ReLU 激活函数的全连接神经网络	利用随机矩阵和代数几何，发现了损失函数与球形旋转玻璃模型的哈密顿量同分布，将损失函数表示成为网络深度的函数	保持参数数量不变，增加网络深度，最优点在参数空间中更加聚集	没有给出临界点是局部极小点还是全局最优点的判据

4. 梯度下降法收敛分析

本节回顾了利用神经切线核、共轭核、Gram 矩阵等对梯度下降算法进行收敛性分析的研究进展.

4.1 收敛性

神经切线核(Neural Tangent Kernel)为研究梯度下降法收敛性提供了一个崭新的分析工具 [33]. Jacot等 [33]发现了隐藏层无限宽的神经网络可等价于高斯过程，提出了神经切线核来描述网络动态训练过程.

$$\Theta_\theta^{(L)} = \sum_{i=1}^{P} \partial_{\theta_p} F^{(L)}(\theta) \otimes \partial_{\theta_p} F^{(L)}(\theta) \tag{6}$$

其中，$\Theta_\theta^{(L)}$ 为层数为 L 的深度网络的神经切线核，$F^{(i)}(\theta)$ 为第 i 层输出，$\partial_{\theta_p} F^{(L)}(\theta)$ 为第 L 层输出对参数的导数，$\otimes$ 代表克罗内克积. Jacot等 [33]证明了损失函数在参数空间的梯度下降等价为在函数空间的神经切线核梯度下降. 神经切线核的正定性保证了梯度下降法收敛，无穷宽网络 l_2 损失函数沿着核矩阵的最大主成分方向收敛最快. Lee等 [34]发现了无穷宽网络损失函数梯度下降动力学特征可用初始化参数的网络输出一阶泰勒展开估计. 该泰勒展开由初始化参数网络的神经切线核和网络输出确定. 证明了神经切线核梯度下降全局收敛. 当网络宽度 n 趋于无穷大时，基于格朗沃尔不等式证明网络实际输出与一阶泰勒展开的距离小于 $O(1/\sqrt{n})$.

Chen等 [35]利用神经切线核随机特征函数 [36]，在初始参数邻域内用一阶泰勒展开近似网络函数，针对ReLU激活函数的二分类神经网络，当网络宽度为输入维数对数的多项式函数时，批量梯度下降算法和随机梯度下降算法具有全局收敛性.

4.2 收敛速度

Sankararaman等 [37]分析了网络结构对小批量随机梯度下降算法收敛速度的影响. 小批量随机梯度下降算法每次从训练集中选取小批量的样本，获得在小批量样本上的梯度后，对网络参

chinaXiv:202112.00002v1

数进行更新. 然而, 所获得的梯度可能负相关, 无法确定损失函数的下降方向, 这一现象称为梯度混淆. 该研究发现, 通过增加网络宽度, 可以降低梯度混淆, 加速算法的收敛. 该研究的结论只能保证小批量随机梯度下降算法可以收敛到一个稳定点, 未给出算法收敛到全局最优点的必要条件.

4.3 收敛性和收敛速度

针对复杂性较低的深层线性神经网络, Arora等 [38]构造了权重矩阵的限定条件, 证明了在恰当的学习率下, 深层线性神经网络损失函数值以一定概率线性收敛到全局最优点.

Du等 [39]基于Gram矩阵, 分析了批量梯度下降算法的收敛性和收敛速度. 神经网络的动力学特征说明了其收敛性依赖于Gram矩阵的最小特征值, 从初始值开始控制Gram矩阵的最小特征值, 可以有效控制其下界; 其次, 当神经网络过参数化时, 权重矩阵会接近其初始值. 根据这两点, 假设Gram矩阵的最小特征值总大于零. 对于全连接的前馈神经网络, 当每层网络中的神经元数量随网络深度的增加呈指数级增长时, 梯度下降算法会在合适步长下以线性速率收敛到零损失函数. 而对于深层残差网络和卷积残差网络, 实现同样的结果只要求每层网络的神经元数量随网络深度呈多项式增长. 上述理论结果仅适用于光滑激活函数或者利普希茨连续的激活函数, 损失函数为l_2, 算法为批量梯度下降算法的情形.

Zou和Gu [40]发现了当随机初始化的权重参数服从高斯分布时, 且当隐藏层的节点个数是输入数据数量、数据点距离以及隐藏层层数的多项式函数时, 利用批量梯度下降算法和随机梯度下降算法在初始权重附近生成微小扰动的序列, ReLU激活函数的二分类网络损失函数具备较好的局部性质, 能保证全局收敛, 收敛速度与假设条件有关.

假设数据不退化, 且各隐藏层的神经元数量是层数和样本数量的多项式函数时, Allen-Zhu等[11]证明了损失函数具备两个重要的性质. 其一, 损失函数不存在鞍点. 只要没有收敛到全局最优点, 损失函数的梯度就一定大于零. 且损失函数值越大, 损失函数梯度的模长就越大. 其二, 损失函数具备半光滑性. 损失函数和其一阶近似之间的距离很小. 基于此, 针对ReLU激活函数的深层网络, Allen-Zhu等证明了批量梯度下降算法和随机梯度下降算法线性收敛到零损失函数. 当激活函数光滑时, 关于隐藏层神经元数量的条件还可以进一步放松, 即当隐藏层神经元数量级非常小时, 上述结论依然成立. Allen-Zhu等 [41]将上述结果推广到了ReLU激活函数的循环神经网络, 同样证明了循环神经网络损失函数不存在鞍点, 且具有半光滑性. 在隐藏层神经元数量是层数与样本数量的多项式函数时, 梯度下降算法使得损失函数线性收敛到零损失函数. 在与Allen-Zhu等 [11]相同的假设条件下, Zou和Gu [42]降低了对网络宽度数量级和优化迭代次数的要求, 获得了更严格的梯度下界和更精确的收敛速度. Daniely [43]利用随机梯度下降算法学习共轭核空间任意函数, 证明了对于层数大于 2 且小于 $\log(n)$ 的全连接、卷积神经网络, 当网络尺寸和迭代次数为输入输出维数的多项式时, 随机梯度下降法能在多项式时间内收敛到零损失函数.

表 2 梯度下降法收敛性分析研究

作者	时间	网络类型	研究方法	结论及适应性	局限性
Daniely [43]	2017	全连接神经网络、卷积神经网络	构建了神经网络和成分核的对偶性质	随机梯度下降法能在多项式时间内收敛到零损失函数	局限于共轭核空间的函数类
Arora等 [38]	2018	深层线性神经网络	提出了权重矩阵的限定约束, 构造了平衡初始化的权重算法	在恰当的学习率下, 损失函数值以一定概率线性收敛到全局最优点	权重矩阵需要满足两个限制条件
Zou等 [40]	2018	ReLu 激活函	证明了高斯随	批量梯度下降算法和随	结论局限于离散型损失函数

chinaXiv:202112.00002v1

		数的二分类深层神经网络	机初始化权重若干性质，在初始权重附近生成微小扰动的序列具备较好的局部性质，能保证全局收敛	机梯度下降算法能保证全局收敛	
Jacot等 [33]	2018	无限宽神经网络	发现了隐藏层无限宽网络等价于高斯过程，提出了神经切线核来描述网络动态训练过程	证明了损失函数在参数空间的梯度下降等价为在函数空间的神经切线核梯度下降，神经切线核正定性保证梯度下降法收敛，无穷宽网络损失函数沿着核矩阵的最大主成分方向收敛最快	为研究梯度下降法收敛性提供了一个崭新的分析工具，目前只适用于宽度趋于无穷的网络
Du等 [39]	2019	全连接神经网络、深层残差网络和卷积残差网络	发现了网络收敛性依赖于Gram 矩阵的最小特征值，过参数化网络的权重矩阵接近其初始值	Gram 矩阵最小特征值大于零保证了梯度下降算法以线性速率收敛到零损失函数	上述理论结果仅适用于光滑激活函数或者利普希茨连续的激活函数
Allen-Zhu等 [11]	2019	ReLU 激活函数的深层网络	证明了损失函数两个重要性质，即不存在鞍点和半光滑性	证明了批量梯度下降算法和随机梯度下降算法线性收敛到零损失函数	给出了宽度的多项式阶估计，但对于实际而言，数量级仍很巨大
Allen-Zhu等 [41]	2019	ReLU 激活函数的循环神经网络	证明了循环神经网络损失函数半光滑且不存在鞍点	梯度下降算法使得损失函数线性收敛到零损失函数	需满足隐藏层神经元数量是层数与样本数量的多项式函数的条件
Zou和Gu [42]	2019	ReLU 激活函数的全连接网络	在初始化权重的特定邻域内，估计损失函数的梯度和步长上界，可证得收敛性与收敛速度，再证明迭代过程中每步参数都在该邻域内	获得了比 Allen-Zhu 等 [11]更严格的梯度下界和更精确的收敛速度	降低了网络宽度和迭代次数的数量级，但较依赖于高斯初始化
Lee等 [34]	2019	无穷宽神经网络	损失函数梯度下降动力学特征可用初始化参数的网络输	神经切线核梯度下降全局收敛	结论适用于无穷宽网络

chinaXiv:202112.00002v1

			出一阶泰勒展开估计		
Chen等 [35]	2019	ReLU 激活函数的二分类神经网络	利用神经切线核随机特征函数，初始参数邻域内用一阶泰勒展开近似网络函数	数据分离条件下，网络宽度为输入维数对数的多项式函数时，批量梯度下降算法和随机梯度下降算法全局收敛	损失函数是交叉熵损失函数，没有考虑连续损失函数
Sankararaman 等 [37]	2020	全连接神经网络、卷积神经网络	发现了无法确定损失函数下降方向的梯度混淆现象，提出了克服梯度混淆以保证收敛速度的条件	通过增加网络宽度，可以降低梯度混淆，加速算法的收敛，保证小批量随机梯度下降算法稳定收敛	只能保证小批量随机梯度下降算法可以收敛到一个稳定点，未能给出算法收敛到全局最优点的必要条件

5. 损失函数地貌可视化

如何借助仅能展示二维或三维信息的可视化方法，直观展示损失函数的高维地貌是挑战性的研究. 本节回顾了基于降维的关键信息选取法，以期获得信息损失较小的损失函数地貌低维展示.

5.1 滤波器归一化法

网络标度不变性是指对权重参数进行缩放，而不影响网络预测结果的性质. 但该性质会妨碍对不同参数下损失函数进行可视化比较. 因此需要对网络参数进行预处理，去除网络标度不变性对可视化的影响.

滤波器归一化法(Filter Normalization)是去除网络标度不变性影响的一种有效方法 [44]. 根据不同滤波器的范数对二维坐标的方向向量进行归一化，使得敏锐度和泛化误差具有更强的相关性，有助于观察损失函数在某些点附近的凸性. 具体而言，随机生成与网络参数相同维数的方向向量，对滤波器进行归一化，获得去除网络标度不变性的方向向量. Li等 [44]利用滤波器归一化方法，对残差网络的损失函数地貌进行了可视化分析，绘制了损失函数在最小点上任意两个随机方向的等高线图. 发现随着网络层数加深，损失函数的地貌结构越发混乱，非凸性增加，在最小点上的测试误差变大，表明极值点泛化能力降低. 一个有趣的发现是残差网络中的跳跃连接结构有效增加了最小点邻域的平坦性，阻止函数地貌向混乱转变.

5.2 主成分分析法

主成分分析法(Principal Component Analysis)通过正交变换将一组变量转换成一组线性不相关的变量，即主成分. 通过提取部分主成分，既能保留原始数据大部分信息，又能起到降维的效果 [45].

Li等 [44]将各次迭代获得的网络参数组成矩阵，对该矩阵使用主成分分析法，确定信息量最大的两个线性不相关的主成分和网络参数在主成分上的投影系数. 将投影系数作为各点的横纵坐标，绘制随机梯度下降法的收敛路径，标明损失函数的等高线，较好地展示下降路径的动态变化.

5.3 多维标度法

多维标度法(Multidimensional Scaling)是一种在低维空间展示距离数据结构的分析技术，可以保证任意一对点在高维和低维空间距离上的相似性 [46]. 根据距离矩阵构造内积矩阵，选择内积矩阵的较大特征值和对应的特征向量来构建向低维空间的投影.

chinaXiv:202112.00002v1

Liao和Poggio [47]以深度卷积神经网络为例，用多维标度法对参数空间进行降维．利用距离矩阵作为相似度矩阵，保证任意两个参数在二维平面上相对距离的一致性，给出了随机梯度下降算法和批量梯度下降算法优化损失函数时的收敛路径．可视化结果表明：损失函数存在多个零损失点，尽管微小扰动会导致不同的收敛路径，但从任意初始点出发均可收敛全局最优点.

5.4 PHATE 方法

PHATE方法(Potential of Heat diffusion for Affinity-based Transition Embedding)是一种基于扩散流形学习的降维方法 [48]．利用局部相似性对局部信息进行编码，再用点对之间的相似性表示扩散概率，通过扩散过程对全局信息进行编码，由此生成距离矩阵，将其视作相似度矩阵，使用多维标度法，获得可视化的低维嵌入.

借助PHATE方法，Horoi等 [49]对残差网络最小点邻域的损失函数曲面的崎岖程度进行了研究，刻画曲面特征与泛化能力之间的关系．Horoi等发现残差网络损失函数存在多个零损失点，但不同点的泛化能力受该点邻域地貌影响，相对于邻域地貌崎岖不平的极小点，邻域地貌较平坦的极小点的泛化误差较小.

6. 挑战和展望

6.1 损失函数地貌分析

1）松弛施加于训练数据和网络参数的约束：为了便于从理论上分析损失函数的地貌，历史文献对训练数据集和网络参数施加了多种限制性条件，如输入数据相互独立 [19, 22, 31]、激活函数可解析 [24, 31, 32]、网络宽度数量级 [24, 29, 31]、损失函数可微 [19]等．应注重离散型损失函数、非光滑激活函数、训练集分布任意等情形下的损失函数地貌特征分析.

2）关注复杂网络结构的损失函数地貌：历史文献重点关注线性网络 [19, 27-30]和全连接网络 [22-24, 26, 31]，利用矩阵分解 [27, 28]和满秩条件 [29, 31]等，得出了临界点最优性的判别条件 [19, 27-29]．应关注卷积、递归等复杂神经网络的最优解特征等.

6.2 梯度下降法收敛性分析

1）松弛网络宽度对全局收敛性分析的限制：历史文献基于网络宽度指数阶或者多项式阶的假设，分析了梯度下降算法的全局收敛性[11, 39-42]．需要关注网络宽度数量级下降时的可优性分析.

2）发展基于网络线性化近似的收敛性理论：神经切线核将梯度下降动力学特征表征为网络输出的一阶泰勒展开，成为分析无限宽网络收敛性的新颖工具 [33, 34, 36, 43]．如何发展非线性网络的线性化近似表达，分析梯度下降法的收敛性，估计泛化误差等有待进一步研究.

3）实现梯度下降算法的加速：加速梯度下降算法以提升收敛速度．如Pascanu等 [50]提出了可以逃离鞍点的快速下降算法，Arora等 [38]提出了平衡初始化算法实现了在一定概率下收敛于全局最优点.

6.3 基于损失函数地貌的泛化误差分析

损失函数地貌可视化可以定性给出地貌崎岖程度与网络泛化能力的关联关系，即零损失点的邻域越崎岖，则在该点的泛化能力越差 [49]．除了研究泛化误差的理论上下界[35, 36]，借助可视化方法解释网络的泛化能力也是一个值得探索的方向.

致谢

感谢中国科学院数学与系统科学研究院汪寿阳研究员、高小山研究员、清华大学金以慧教授、王凌教授、北京大学黄季焜教授、国家呼吸系统疾病临床研究中心何建行教授的建议和帮助.

参考文献

[1] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.

chinaXiv:202112.00002v1

[2] Schmidhuber, Jürgen. Deep Learning in Neural Networks: An Overview[J]. Neural Netw, 2015, 61: 85-117.
[3] Liu B, Chi W, Li X, et al. Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect[J]. Journal of cancer research and clinical oncology, 2020, 146(1): 153-185.
[4] Yang Y, Feng X, Chi W, et al. Deep learning aided decision support for pulmonary nodules diagnosing: a review[J]. Journal of thoracic disease, 2018, 10(Suppl 7): S867.
[5] Hochreiter S. Untersuchungen zu dynamischen neuronalen Netzen[J]. Diploma, Technische Universität München, 1991, 91(1).
[6] Fukushima K, Miyake S, Ito T. Neocognitron: A neural network model for a mechanism of visual pattern recognition[J]. IEEE transactions on systems, man, and cybernetics, 1983, (5): 826-834.
[7] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[J]. arXiv preprint arXiv:1611.03530, 2016.
[8] Goodfellow I J, Vinyals O, Saxe A M. Qualitatively characterizing neural network optimization problems[J]. arXiv preprint arXiv:1412.6544, 2014.
[9] Telgarsky M. Representation benefits of deep feedforward networks[J]. arXiv preprint arXiv:1509.08101, 2015.
[10]Li D, Ding T, Sun R. On the benefit of width for neural networks: Disappearance of bad basins[J]. arXiv, 2018: arXiv: 1812.11039.
[11]Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization[C]// International Conference on Machine Learning, 2019: 242-252.
[12]Csáji B C. Approximation with artificial neural networks[J]. Faculty of Sciences, Etvs Lornd University, Hungary, 2001, 24(48): 7.
[13]Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural networks, 1989, 2(5): 359-366.
[14]Lu Z, Pu H, Wang F, et al. The expressive power of neural networks: A view from the width[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6232-6240.
[15]Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[16]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.
[17]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems, 2012: 1097-1105.
[18]Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[19]Kawaguchi K. Deep learning without poor local minima[C]// Advances in Neural Information Processing Systems, 2016: 586-594.
[20]Dauphin Y N, Pascanu R, Gulcehre C, et al. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization[C]// Advances in Neural Information Processing Systems, 2014: 2933-2941.
[21]Ballard A J, Das R, Martiniani S, et al. Energy landscapes for machine learning[J]. Physical Chemistry Chemical Physics, 2017, 19(20): 12585-12603.
[22]Choromanska A, Henaff M, Mathieu M, et al. The loss surfaces of multilayer networks[C]//

chinaXiv:202112.00002v1

Artificial intelligence and statistics: PMLR, 2015: 192-204.
[23]Becker S, Zhang Y. Geometry of energy landscapes and the optimizability of deep neural networks[J]. Physical review letters, 2020, 124(10): 108301.
[24]Cooper Y. The loss landscape of overparameterized neural networks[J]. arXiv preprint arXiv:1804.10200, 2018.
[25]Garipov T, Izmailov P, Podoprikhin D, et al. Loss surfaces, mode connectivity, and fast ensembling of dnns[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 8803-8812.
[26]Nguyen Q. On connected sublevel sets in deep learning[C]// International Conference on Machine Learning: PMLR, 2019: 4790-4799.
[27]Lu H, Kawaguchi K. Depth creates no bad local minima[J]. arXiv preprint arXiv:1702.08580, 2017.
[28]Zhou Y, Liang Y. Critical points of neural networks: Analytical forms and landscape properties[J]. arXiv preprint arXiv:1710.11205, 2017.
[29]Yun C, Sra S, Jadbabaie A. Global optimality conditions for deep neural networks[J]. arXiv preprint arXiv:1707.02444, 2017.
[30]Hardt M, Ma T. Identity matters in deep learning[J]. arXiv preprint arXiv:1611.04231, 2016.
[31]Nguyen Q, Hein M. The loss surface of deep and wide neural networks[C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70: JMLR. org, 2017: 2603-2612.
[32]Nguyen Q, Hein M. Optimization landscape and expressivity of deep cnns[C]// International conference on machine learning: PMLR, 2018: 3730-3739.
[33]Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks[J]. arXiv preprint arXiv:1806.07572, 2018.
[34]Lee J, Xiao L, Schoenholz S, et al. Wide neural networks of any depth evolve as linear models under gradient descent[J]. Advances in Neural Information Processing Systems, 2019, 32: 8572-8583.
[35]Chen Z, Cao Y, Zou D, et al. How much over-parameterization is sufficient to learn deep relu networks?[J]. arXiv preprint arXiv:1911.12360, 2019.
[36]Cao Y, Gu Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32: 10836-10846.
[37]Sankararaman K A, De S, Xu Z, et al. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent[C]// International Conference on Machine Learning: PMLR, 2020: 8469-8479.
[38]Arora S, Cohen N, Golowich N, et al. A convergence analysis of gradient descent for deep linear neural networks[J]. arXiv preprint arXiv:1810.02281, 2018.
[39]Du S, Lee J, Li H, et al. Gradient descent finds global minima of deep neural networks[C]// International Conference on Machine Learning: PMLR, 2019: 1675-1685.
[40]Zou D, Cao Y, Zhou D, et al. Stochastic Gradient Descent Optimizes Over-parameterized Deep ReLU Networks. arXiv e-prints, art[J]. arXiv preprint arXiv:1811.08888, 2018.
[41]Allen-Zhu Z, Li Y, Song Z. On the Convergence Rate of Training Recurrent Neural Networks[J]. Advances in Neural Information Processing Systems, 2019, 32: 6676-6688.
[42]Zou D, Gu Q. An improved analysis of training over-parameterized deep neural networks[C]// Advances in Neural Information Processing Systems, 2019: 2055-2064.
[43]Daniely A. SGD learns the conjugate kernel class of the network[J]. arXiv preprint
chinaXiv:202112.00002v1

arXiv:1702.08503, 2017.
[44]Li H, Xu Z, Taylor G, et al. Visualizing the loss landscape of neural nets[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 6391-6401.
[45]Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.
[46]Buja A, Swayne D F, Littman M L, et al. Data visualization with multidimensional scaling[J]. Journal of Computational and Graphical Statistics, 2008, 17(2): 444-472.
[47]Liao Q, Poggio T. Theory II: Landscape of the empirical risk in deep learning[J]. arXiv preprint arXiv:1703.09833, 2017.
[48]Moon K R, Van Dijk D, Wang Z, et al. Visualizing structure and transitions in high-dimensional biological data[J]. Nature biotechnology, 2019, 37(12): 1482-1492.
[49]Horoi S, Huang J, Wolf G, et al. Visualizing high-dimensional trajectories on the loss-landscape of ANNs[J]. arXiv preprint arXiv:2102.00485, 2021.
[50]Pascanu R, Dauphin Y N, Ganguli S, et al. On the saddle point problem for non-convex optimization[J]. arXiv preprint arXiv:1405.4604, 2014.

chinaXiv:202112.00002v1